Volume 1, number 10

What is a Cox model?

Sponsored by an educational grant from Aventis Pharma

Stephen J Walters
BSc MSc CStat
Statistician,
School of Health and Related Research (ScHARR),
University of Sheffield

- A **Cox model** is a well-recognised statistical technique for exploring the relationship between the survival of a patient and several explanatory variables.

- **Survival analysis** is concerned with studying the time between entry to a study and a subsequent event (such as death). **Censored** survival times occur if the event of interest does not occur for a patient during the study period.

- A Cox model provides an **estimate of the treatment effect on survival** after adjustment for other explanatory variables. It allows us to estimate the hazard (or risk) of death, or other event of interest, for individuals, given their prognostic variables.

- Even if the treatment groups are similar with respect to the variables known to effect survival, using the Cox model with these prognostic variables may produce a more precise estimate of the treatment effect (for example, by narrowing the confidence interval).

- Interpreting a Cox model involves examining the coefficients for each explanatory variable. A **positive regression coefficient** for an explanatory variable means that the **hazard is higher**, and thus the prognosis worse, for higher values. Conversely, a **negative regression coefficient** implies a **better prognosis** for patients with higher values of that variable.

Abbreviated prescribing information is on page 7

What is the purpose of the Cox model?

The Cox model is based on a modelling approach to the analysis of survival data. The purpose of the model is to simultaneously explore the effects of several variables on survival.

The Cox model is a well-recognised statistical technique for analysing survival data. When it is used to analyse the survival of patients in a clinical trial the model allows us to isolate the effects of treatment from the effects of other variables. The model can also be used *a priori*, if it is known that there are other variables besides treatment that influence patient survival and these variables cannot easily be controlled for in a clinical trial. Using the model may improve the estimate of treatment effect by narrowing the confidence interval.

Survival times now often refer to the development of a particular symptom or to relapse after remission of a disease, as well as to the time to death. In the amyotrophic lateral sclerosis example (described later) the survival time was counted as the time either to death or to undergoing a tracheostomy.

Why are survival times censored?

A significant feature of survival times is that the event of interest is very rarely observed in all subjects. For example, in a study to compare the survival of patients having different types of surgery for liver cancer, although the patients may be followed up for several years, there will be some patients who are still alive at the end of the study. We do not, therefore, know what their survival time is from surgery, only that it will be longer than their time in the study. Such survival times are termed **censored**, to indicate that the period of observation was cut off before the event of interest occurred.

From a set of observed survival times (including censored times) in a sample of individuals, we can estimate the proportion of the population of such people who would survive a given length of time under the same circumstances. This method is called the product limit or **Kaplan–Meier method**. The method allows a table and a graph to be produced; these are referred to as the life table and survival curve respectively.

Kaplan–Meier estimate of the survivor function

To determine the Kaplan–Meier estimate of the **survivor function**, taking hypothetical data on 18 patients following surgery for liver cancer, a series of time intervals is formed. Each of these intervals is constructed to be such that one observed death is contained in the interval, and the time of this death is taken to occur at the start of the interval.

Table 1 shows the survival times arranged in ascending order (column A). Some survival times (*) are censored. The number of patients who are alive just before 10 weeks is 18 (column B). Since one patient dies at 10 weeks (column C), the probability of dying by 10 weeks is 1/18 = 0.0556. So the corresponding probability of surviving up to 10 weeks is 1 minus the probability of dying (column E), or 0.9444. The next two time intervals contain censored data, so the probability of surviving in these time intervals is 1. The cumulative probability of surviving up to 19 weeks, then, is the probability of surviving at 19 weeks and surviving throughout all the preceding time intervals – ie, 0.9444 x 1.0000 x 1.0000 x 0.9333 = 0.8815 (column F). This is the Kaplan–Meier estimate of the survivor function.

Sometimes there are censored survival times which occur at the same time as deaths. The censored survival time is then taken to occur immediately after the death time when calculating the survivor function.

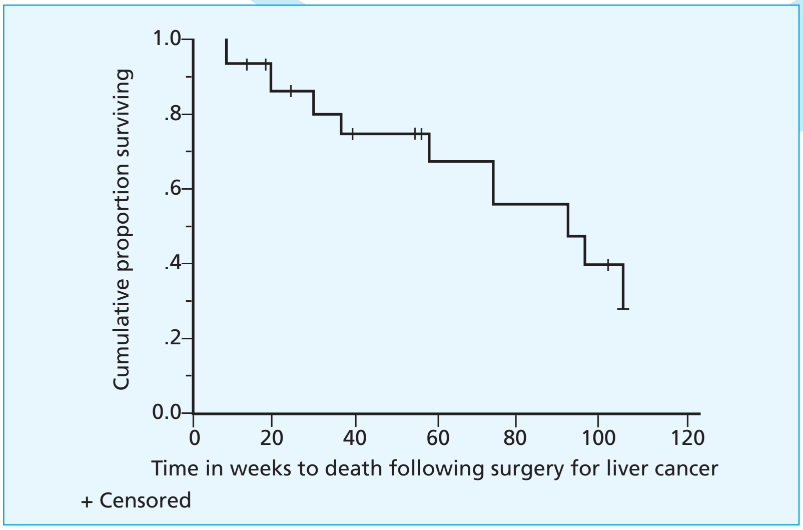
A plot of the Kaplan–Meier estimate of the survivor function (Figure 1) is a step function, in which the estimated survival probabilities are constant between adjacent death times and only decrease at each death.

Table 1. Calculation of Kaplan–Meier estimate of the survivor function

A Time (weeks)	B Number at risk at start of study	C Number of deaths	D Number censored	E Proportion surviving until end of week	F Cumulative proportion surviving
10	18	1	0	1 – 1/18 = 0.9444	0.9444
13*	17	0	1	1 – 0/17 = 1.0000	0.9444
18*	16	0	1	1 – 0/16 = 1.0000	0.9444
19	15	1	0	1 – 1/15 = 0.9333	0.8815
23*	14	0	0	1 – 0/17 = 1.0000	0.8815
30	13	1	0	1 – 1/13 = 0.9230	0.8137
36	12	1	0	1 – 1/12 = 0.9167	0.7459
38*	11	0	1	1 – 0/17 = 1.0000	0.7459
54*	10	0	1	1 – 0/17 = 1.0000	0.7459
56*	9	0	1	1 – 0/17 = 1.0000	0.7459
59	8	1	0	1 – 1/8 = 0.8750	0.6526
75	7	1	0	1 – 1/7 = 0.8571	0.5594
93	6	1	0	1 – 1/6 = 0.8300	0.4662
97	5	1	0	1 – 1/5 = 0.8000	0.3729
104*	4	0	1	1 – 0/4 = 1.0000	0.3729
107	3	1	0	1 – 1/3 = 0.6667	0.2486
107*/107*	2	0	2	1 – 0/2 = 1.0000	0.2486

An important part of survival analysis is to produce a plot of the survival curves for each group of interest. However, the comparison of the survival curves for two groups should be based on a formal statistical test called the logrank test, and not on visual impressions.[1]

Figure 1. Kaplan–Meier survival curve



Modelling survival – the Cox regression model

The logrank test cannot be used to explore (and adjust for) the effects of several variables, such as age and disease duration, known to effect survival. Adjustment for variables that are known to affect survival may improve the precision with which we can estimate the treatment effect.

The regression method introduced by Cox is used to investigate several variables at a time.[2] It is also known as **proportional hazards regression analysis**.

Cox's method does not assume a particular distribution for the survival times, but rather assumes that the effects of the different variables on survival are constant over time and are additive in a particular scale.

The actual method is much too complex for detailed discussion here. This publication is intended to give an introduction to the method, and should be of use in the understanding and interpretation of the results of such analyses.

What is a hazard function?

The **hazard function** is the probability that an individual will experience an event (for example, death) within a small time interval, given that the individual has survived up to the beginning of the interval. It can therefore be interpreted as the risk of dying at time *t*.

The hazard function (denoted by h(*t*)) can be estimated using the following equation:

$$h(t) = \frac{\text{number of individuals experiencing an event in interval beginning at } t}{(\text{number of individuals surviving at time } t) \text{ x } (\text{interval width})}$$

What is regression?

If we want to describe the relationship between the values of two or more variables, we can use a statistical technique called **regression**. If we have observed the values of two variables, say X (age of children) and Y (height of children), we can perform a regression of Y on X. We are investigating the relationship between a **dependent variable** (the height of children) based on the **explanatory variable** (the age of children).

When more than one explanatory (X) variable needs to be taken into account the method is known as **multiple regression**. Cox's method is similar to multiple regression analysis, except that it allows us to take more than one explanatory variable into account at any one time (for example, age, disease duration, bulbar signs at entry, location of trial). We can express the hazard or risk of dying at time *t* as:

$$h(t) = h_0(t) \text{ x } \exp(b_{age}.age + b_{duration}.duration + \ldots + b_{location}.location).$$

Taking natural logarithms of both sides:

$$\ln h(t) = \ln h_0(t) + b_{age}.age + b_{duration}.duration + \ldots + b_{location}.location.$$

The quantity $h_0(t)$ is the baseline or underlying hazard function, and corresponds to the probability of dying (or reaching an event) when all the explanatory variables are zero. The baseline hazard function is analogous to the intercept in ordinary regression (since $\exp^0 = 1$).

The regression coefficients b_{age} to $b_{location}$ give the proportional change that can be expected in the hazard, related to changes in the explanatory variables. They are estimated by a complex statistical method called maximum likelihood,[3] using an appropriate computer program (for example, SAS or SPSS).

The assumption of a constant relationship between the dependent variable and the explanatory variables is called **proportional hazards**. The conclusions reached using this method should be tested.[3]

Interpretation of the model

As mentioned above, the Cox model must be fitted using an appropriate computer program. The final model from a Cox regression analysis will yield an equation for the hazard as a function of several explanatory variables (including treatment). So how do we interpret the results? This is illustrated by the following example.

Cox regression analysis was carried out on the data from a randomised trial comparing riluzole and placebo in the treatment of amyotrophic lateral sclerosis (ALS).[4]

ALS is a progressive, fatal neurodegenerative disorder, characterised by progressive loss of motor neurones. In this trial, 959 patients with clinically probable or definite ALS of less than five years duration were randomly assigned to one of four treatment groups: placebo, 50 mg, 100 mg or 200 mg of riluzole (Rilutek®) daily. The aim of this multicentre, double blind study was to assess the efficacy of riluzole at different doses. The primary outcome was survival without a tracheostomy. Patients were followed up for 18 months from randomisation.[4]

The Cox model included 10 variables as independent prognostic factors, plus a further three treatment variables (see Table 2). Each variable was statistically significant at the 5% level at least.[4] An approximate test of significance for each variable is carried out by dividing the regression estimate (b) by its standard error (SE(b)), and comparing the result with the standard normal distribution. Values of this ratio greater than 1.96 will be statistically significant at the 5% level. The Cox model is shown in Table 2.

The first feature to note in such a table is the sign of the regression coefficients. A positive sign means that the hazard (risk of a tracheostomy or death) is higher, and thus the prognosis worse, for subjects with higher values of that variable. Thus from Table 2, older values of age and increased CGI severity are associated with poorer survival, while people who had already survived the disease for some years or were heavier on entry to the study had better prognosis. The binary variable 'bulbar signs at entry' shows a worse prognosis for subjects with bulbar signs at entry. The three treatment variables show better prognosis for subjects treated with riluzole rather than placebo.

An individual regression coefficient is interpreted quite easily. From Table 2, the estimated hazard with 100 mg of riluzole is exp(-0.43) = 0.65 of that with placebo – ie, a 35% decrease in the risk of death or tracheostomy. At 18 months, the 50 mg, 100 mg and 200 mg riluzole doses decreased the risk of death or tracheostomy by 24%, 35% and 39% respectively, after adjustment for the other explanatory variables in the model.

For explanatory variables which are continuous (for example, age, disease duration, weight), the regression coefficient refers to the increase in log hazard for an increase of 1 in the value of the covariate.

Table 2. Cox regression model adapted from the data from the ALS trial of riluzole versus placebo (n = 959)

Variable	Regression coefficient (b)	Standard error SE(b)	P	e^b Relative risk* (95% CI)
Treatment				
50 mg versus placebo	-0.27	0.15	0.04	0.76 (0.59–0.99)
100 mg versus placebo	-0.43	0.16	0.002	0.65 (0.50–0.85)
200 mg versus placebo	-0.49	0.16	0.0004	0.61 (0.47–0.80)
Demographic and clinical baseline variables				
Age (per 10 years)	0.39	0.05	0.0001	1.48 (1.33–1.63)
Disease duration (years)	-0.40	0.05	0.0001	0.67 (0.61–0.74)
Bulbar signs at entry (yes/no)	0.28	0.15	0.03	1.32 (1.02–1.71)
Weight (per 5 kg)	-0.06	0.02	0.005	0.94 (0.90–0.98)
Vital capacity (per 10% of normal)	-0.25	0.03	0.0001	0.78 (0.74–0.83)
VAS stiffness (per 10 mm)	-0.05	0.02	0.02	0.95 (0.92–0.99)
VAS tiredness (per 10 mm)	0.09	0.02	0.0001	1.09 (1.05–1.14)
Muscle testing (per 5 points)	-0.11	0.02	0.0001	0.90 (0.87–0.94)
CGI severity (per point score)	0.14	0.08	0.04	1.15 (1.01–1.32)
Location				
France and Belgium versus rest of Europe	-0.39	0.14	0.002	0.68 (0.54–0.87)
North America versus France and Belgium	-0.37	0.18	0.02	0.69 (0.51–0.94)

*Risk of death or tracheostomy (with 95% CI) according to treatment assignment and prognostic variables.

VAS = visual analogue scale (range 0–100 mm);
CGI = clinical global impression;
Muscle testing (range 5–110).

Thus at 18 months (Table 2), the estimated hazard or risk of death increases by exp(0.39) = 1.48 times if a patient is 10 years older, after adjustment for the effects of the other variables in the model.

In this Cox model, the country group where the patient was treated was also a significant predictor of survival. Patients in France or Belgium had a higher risk of death/tracheostomy than did patients in other European countries or North America.

The individual overall effect on the survival probability, however, cannot be described simply as it depends on the patient's values of the other variables in the model.

However, the authors were able to conclude that riluzole treatment of ALS had a positive and dose-dependent effect on tracheostomy-free survival, even after adjustment for prognostic factors.

Box 1. Glossary of terms

CI or **confidence interval**. A range of values, calculated from the sample of observations that are believed, with a particular probability, to contain the true parameter value. A 95% confidence interval implies that if the estimation process was repeated again and again, 95% of the calculated intervals would be expected to contain the true parameter value. Note that the stated probability level refers to the properties of the interval and not to the parameter itself.

Logarithms. Logarithms are mainly used in statistics to transform a set of observations to values with a more convenient distribution. The natural logarithm (**$\log_e x$** or **ln x**) of a quantity x is the value such that $x = e^y$. Here e is the constant 2.718281…. The log of 1 is 0 and the log of 0 is minus infinity. **e^x** or **exp(x)** is the exponential function, denoting the inverse procedure to that of taking logarithms.

SE or **se.** The standard error of a sample mean or some other estimated statistic (for example, regression coefficient). It is the measure of the uncertainty of such an estimate and is used to derive a confidence interval for the population value. The notation SE(b) means the standard error of b.

P. The probability value, or significance level, from a hypothesis test. P is the probability of the data (or some other more extreme data) arising by chance when the null hypothesis is true.

References

1. Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall, 1991: 365–396.
2. Cox DR. Regression models and life tables. *J Roy Statist Soc B* 1972; **34:** 187–220.
3. Collett D. *Modelling Survival Data in Medical Research*. London: Chapman & Hall, 1994.
4. Lacombiez L, Bensimon G, Leigh PN *et al*. Dose-ranging study of riluzole in amyotrophic lateral sclerosis. *Lancet* 1996; **347:** 1425–1431.

Further reading

Chapter 13 of Altman[1] provides a good introduction to survival analysis, the logrank test and the Cox regression model. A more detailed technical discussion of survival analysis and Cox regression is given by Collett.[3]

Abbreviated prescribing information: Rilutek®

Presentation: Rilutek Tablets contain riluzole 50mg. **Indications:** Riluzole is indicated to extend life or the time to mechanical ventilation for patients with amyotrophic lateral sclerosis (ALS). Clinical trials have demonstrated that Rilutek extends survival for patients with ALS. There is no evidence that riluzole exerts a therapeutic effect on motor function, lung function, fasciculations, muscle strength or motor symptoms. Riluzole has not been shown to be effective in the late stages of ALS. The safety and efficacy of riluzole has only been studied in ALS. **Dosage and administration:** Adults and Elderly: One 50mg tablet bd; Children: Not recommended; Renal impairment: Not recommended; Hepatic impairment: See warnings and precautions. **Contra-indications:** Severe hypersensitivity to riluzole. Patients with hepatic disease where baseline transaminases are greater than 3 times ULN. Pregnancy, breast feeding. **Warnings and Precautions:** Prescribe with care in patients with history of abnormal liver function or patients with increased transaminase, bilirubin and/or GGT levels. Measure serum transaminases regularly during initiation of treatment with riluzole and frequently in patients who develop elevated ALT levels during treatment. Treatment should be discontinued if ALT level increases to 5 times ULN. Discontinue riluzole in the presence of neutropenia. Any febrile illness must be reported to the physician. Do not drive or use machines if vertigo or dizziness are experienced. **Interactions:** *In vitro* data suggests CYP 1A2 as the primary isozyme in the oxidative metabolism of riluzole; inhibitors or inducers of CYP 1A2 may affect the elimination of riluzole. **Pregnancy and lactation:** Contra-indicated. **Side effects:** Asthenia, nausea and elevations in LFT's are the most frequent events seen. Less frequent events include pain, vomiting, dizziness, tachycardia, somnolence and circumoral paraesthesia. **Legal Category:** POM. **Package Quantities and Basic NHS Price:** Each box of Rilutek Tablets contains 4 blisters of 14 tablets; £286.00.

Marketing Authorisation Number: Rilutek tablets 50mg EU/1/96/010/001.

Full Prescribing Information and further information is available on request from Aventis Pharma Limited, 50 Kings Hill Avenue, Kings Hill, West Malling, Kent. ME19 4AH. **Date of preparation:** November 2000.

Volume 1, number 10

What is a Cox model?

This publication, along with the others in the series, is available on the internet at www.evidence-based-medicine.co.uk

Sponsored by an educational grant from *Aventis Pharma*

Any enquiries please contact:
Team Assistant to the Health Economics Unit
Aventis House
Kings Hill
West Malling
Kent
ME19 4AH
Tel: 01732 584 254
Fax: 01732 584 029

Published by Hayward Medical Communications, a division of Hayward Group plc.

RIL 73929
Date of preparation: May 2001